

Inferring disease causing genes and their pathways: A mathematical perspective

Jeethu V. Devasia

Department of Computer Science and Engineering
National Institute of Technology Calicut,
India - 673 601
Email: jeethu_p130021cs @ nitc.ac.in

Priya Chandran

Department of Computer Science and Engineering,
National Institute of Technology Calicut,
India - 673 601
Email: priya @ nitc.ac.in

Abstract—Background and Objective: A system level view of cellular processes for human and several organisms can be captured by analyzing molecular interaction networks. A molecular interaction network formed of differentially expressed genes and their interactions helps to understand key players behind disease development. So, if the functions of these genes are blocked by altering their interactions, it would have a great impact in controlling the disease. Due to this promising consequence, the problem of *inferring disease causing genes and their pathways* has attained a crucial position in computational biology research. However, considering the huge size of interaction networks, executing computations can be costly. Review of literatures shows that the methods proposed for finding the set of disease causing genes could be assessed in terms of their accuracy which a perfect algorithm would find. Along with accuracy, the time complexity of the method is also important, as high time complexities would limit the number of pathways that could be found within a pragmatic time interval.

Methods and Results: Here, the problem has been tackled by integrating graph theoretical approaches with an approximation algorithm. The problem of *inferring disease causing genes and their pathways* has been transformed to a graph theoretical problem. Graph pruning techniques have been applied to get the results in practical time. Then, randomized rounding, an efficient approach to design an approximation algorithm, has been applied to fetch the most relevant causal genes and pathways. Experimentation on multiple benchmark datasets has been demonstrated more accurate and computationally time efficient results than existing algorithms. Also, biological relevance of these results has been analyzed.

Conclusions: Based on computational approaches on biological data, the sets of disease causing genes and corresponding pathways are identified for multiple disease cases. The proposed approach would have a remarkable contribution in areas like drug development and gene therapy, if we could recognize these results biologically too.

Index Terms—Molecular interaction Network; Causal genes; Dysregulated pathway; Graph pruning; Approximation algorithm; Randomized rounding

I. INTRODUCTION

Cellular processes are mainly governed by the co-action of biomolecules. For example, a particular protein function can be understood by mapping protein-protein interactions. These biological communications can be represented using molecular interaction networks. At present, molecular interaction networks of various organisms are available [1]. We can exploit them for diverse aims such as discovering

disease causing genes and related pathways. From the computational perspective, computing information flows in a complex model is expensive as the input sizes are large and the analyses typically have very high time complexities. Therefore, proposing algorithms for effective computation on the networks would have a remarkable impact on the knowledge to be gained from such networks. The problem of *inferring disease causing genes and dysregulated pathways* is of prime importance and huge academic and industrial interest, because, it is potentially very useful for comprehending the underlying system of complex diseases and suggesting prospective drug targets. An algorithm that augments graph theoretical approaches with approximation for *inferring causal genes and dysregulated pathways* is addressed in this paper. The proposed method incorporates gene expression value because of its potentiality in predicting diseases. High-risk genes are more correlated with each other than the genes with lower risk and vice versa [2]. An experimental analysis of the state of the art related works together with the proposed method is also given in this paper. Related works are based on Random walk based approach, Electric circuit model with Expression Quantitative Loci (eQTL) analysis, Electric circuit model with multiple sources and sinks, Fast iterative matrix inversion and Approximation algorithm based on Randomized rounding [3], [4], [5], [6], [1], [7], [8], [9].

Proposed by Tu et al., the random walk approach has shown a significant impact in the problem of identifying *causal genes* and the underlying *pathways* [1], [3], [4], [6]. Based on the Pearson's correlation coefficient of *gene expression values* of genes, transition probability is defined. Starting from a source node, random walker moves to a node that is qualified as an unvisited and highest transition probability bearing node among all the neighbors of the current node. This process is repeated until it visits the destination node or further movement is not possible. *Candidate causal gene* that has the largest number of visit times or that has the largest value of probability of being a *causal gene* is taken as the *causal gene*, g_c . Identification of *pathway* is done by tracing a path from g_c to the corresponding source node by selecting intermediate nodes as the most visited ones. According to

Suthram et al., this approach results in relatively short walks with the requirement of multiple iterations for better results [6]. They proposed a new approach based on Electric circuit model which is analogous to Random walk based approach [6], [1], [5], [3]. Considering a network of protein-protein interactions and Transcription Factor (TF)-DNA interactions as an electric circuit, conductance of each edge (u, v) is set based on the correlation of *expression values* of u and v with the *target gene*. After solving the electric circuit using Kirchhoff's and Ohm's Laws to get the current through each node and edge, the *causal gene* is taken as the gene with highest value of current flow. *Pathway* is the shortest route between a *target gene* and *causal gene* with the highest total sum of currents across its interactions such that each edge corresponds to an interaction. All such paths together give the *textitpathways* for the entire network.

Based on the above approaches, Y. A. Kim et al. suggested an electric circuit based approach with multiple sources and sinks [1], [7]. Conductance is defined as the average of the absolute value of the Pearson correlation coefficient between *gene expression values* of *target gene* and genes at the end-points of each link. A system of linear equations based on Ohm's law and Kirchhoff's law is solved to find the voltages of links and thereby to calculate the current value. *Causal genes* are taken as the genes with significant amount of current flow. *Pathway* is the shortest paths in the collection of all maximum current paths for each pair of source and sink. Focusing on the faster computation of voltages of nodes suggested in the previous approach, a fourth order iterative method for fast iterative matrix inversion was proposed in [8]. Following the calculation of voltages of nodes, causal genes and dysregulated pathways are identified as in citeref:yoo, [7].

A new approximation algorithm was proposed in [9], based on electric circuit approach to fetch *causal genes* and corresponding *pathways*. Collection of all distinct paths in a reduced network obtained by thresholding the edge-weights is computed. Then, randomized rounding is applied to get the path with maximum current flow which is taken as the *dysregulated pathway* for the corresponding *target* and *causal genes*. The members of the *pathways* are taken as *causal genes*.

Here, we infer that the reported works in literature for finding the set of disease causing genes could be evaluated in terms of their accuracy, i.e, the closeness of the set found to the actual set and execution time. During the assessment, it has been observed that in most of the cases, execution time rises with accuracy.

II. MATERIALS AND METHODS

A. Selection of target genes and candidate causal genes

Gene expression data of Breast cancer and Lung cancer of the species *Homosapiens* (GSE44024 and GSE43459) and Pancreatic cancer of the species *Rattus norvegicus* (GSE22537) have been utilized for experimentation. Data have been collected from the National Center for Biotechnology Information (NCBI) sponsored Gene Expression Omnibus data

repository [9]. Platform details and sample information of the datasets are given in Table ??.

Disease	GEO Accession	Sample count (case / control)	Platform
Breast cancer	GSE44024	4 (2 / 2)	GPL571 (Affymetrix Human Genome U133A 2.0 Array)
Lung cancer	GSE43459	6 (3 / 3)	GPL6244 (Affymetrix Human Gene 1.0 ST Array)
Pancreatic cancer	GSE22537	18 (9 / 9)	GPL1355 (Affymetrix Rat Genome 230 2.0 Array)

TABLE I: Dataset details

The following steps have been done for each dataset separately. The initial data have been normalized using Robust Multi-array Average method to remove any noise due to non-biological factors [10]. Then, the genes having statistical significance have been selected using *t*-test with equal variance and 2-tailed followed by the calculation of *p*-value. Then, significant *q*-values have been computed and genes with *q*-value < 0.05 have been selected. This set of *differentially expressed genes* has been taken as the *candidate causal genes* for Breast cancer and Pancreatic cancer. Considering the data size, first 100 genes after sorting *p*-value in ascending order following the filtering based on *q*-value, have been selected as *candidate causal genes* for Lung cancer. Gene interaction network of molecular interactions has been downloaded from BioGRID database for each *candidate causal gene*. The fetched network data have been filtered to select only the genes that are *differentially expressed*. Genes that are linked with transcription factors in the network have been considered as *target genes*. The set of *target genes* has been considered as the set source nodes. The set of genes apart from the *target genes* has been considered as the set of sink nodes.

B. Selection of benchmark data

Data from NCBI sponsored Gene database, Aceview and Uniprot database and Cosmic database have been curated as benchmark data for Breast cancer and Lung cancer of the species, *Homosapiens*. For this, the set of associated genes corresponding to a particular disease case and species has been fetched from these databases. These fetched genes after removing duplicates have been taken as the benchmark data for each disease. Similarly, data from NCBI sponsored Gene database, Aceview, Uniprot database and literatures from NCBI, Nature, Nucleic Acid Research have been compiled as benchmark data for the species *Rattus norvegicus* [11], [12].

C. Formulation of the algorithm

1) *Biological Background*: Each gene g is expressed to a particular level during the process of producing the ultimate products called proteins and this level can be measured as a numerical quantity, known as expression value, $e(g)$ [13],

[14], [15], [16]. Also, each gene interacts with several genes resulting in certain phenotypes [17]. With the advances in the area of computational biology, molecular interactions can be represented as a network by designating genes/proteins as nodes and edges as associations between end nodes [18], [19]. These associations have different aspects like physical interactions, membership in the same pathway, co-expression and literature co-occurrence [20], [21]. It reveals the fact that two nodes may be linked together in different cases, resulting in a network with multiple edges [22], [23], [24]. Also, molecular interaction network consists of certain directional links such as protein-protein interactions (bidirectional), TF-DNA interactions (directional) and Phosphorylation events (directional) [1], [7], [25], [9].

Expression levels or values of genes help to make distinction between healthy and disease cases in view of the fact that there is an increase or decrease in expression values of some genes in many diseases from that of healthy individuals [26], [27], [28]. When a particular gene's expression value changes between two groups of healthy and affected individuals, then the gene is said to be *differentially expressed* [27], [16], [29]. Differentially expressed genes may lead to a disease state or to a beneficial state.

2) *Definitions*: The set of genes that gives rise to a particular disease state is termed as *causal genes*. The set of probable genes of a certain disease is termed as *candidate causal genes*. The set of *candidate causal genes* that are bound by Transcription Factors is referred to as *target genes* [1], [7], [3].

As described in Section II-C1, *candidate causal genes* are connected together in a gene/protein network due to different aspects. A disease state may be developed by the interference of a *target gene* and a *candidate causal gene* in the normal biological functions of a cell and this relationship between a *target gene* and a *candidate causal gene* is known as dependence [28]. The dependence between a target gene, t and a candidate causal gene, c is computed as

$$\delta(t, c) = P(e(t), e(c)) \quad (1)$$

where, $P(e(t), e(c))$ is the *Pearson's correlation coefficient* of the *expression values* of genes t and c .

Genes with low values of δ and high values of δ are crucial in molecular interactions [30]; so we consider the absolute value of δ [4]. Similarly, in [1] and [7], the edge weight for the edge in the molecular interaction network is taken as the average of the absolute value of the *dependence* of genes (end nodes of the edge) with the *target gene*. The edge weights can be used as a measure of inferring *causal genes* and corresponding *pathways*.

D. Formal notations and definitions

Let G_t be the set of *target genes*, G_{cc} be the set of *candidate causal genes* and G_c be the set of *causal genes*.

The basic problem is to find G_c , the set of *causal genes*, where $G_c \subseteq G_{cc}$ and G_{cc} and G_t are known.

The fundamental decision problem here is

Does $g_{cc} \in G_c$?

where $g_{cc} \in G_{cc}$ [or, is a candidate causal gene, a causal gene].

Let us recall that $g_{cc} \in G_c$ if it has a role in disease, and would be determined by the interactions it has with the *target genes*, $g_t \in G_t$, and other *candidate genes* which have a role in the disease.

For each *target gene* g_t , a weighted graph (network) is defined with g_t as the source, members in G_{cc} as the other (non-source) nodes, and molecular interactions as the edges. Edge weights reflect the role of the genes represented by the end nodes in causing the disease. It may be realized at this point that there would be $|G_t|$ such graphs, each containing $|G_{cc}| + 1$ nodes. Let (\mathcal{G}_t, w_t) be such a network for *target gene*, g_t .

$$\mathcal{G}_t.V = \{g_t\} \cup G_{cc}$$

$\mathcal{G}_t.E$ is defined by molecular interactions, as described next. Let $c(g_{cc}, g_t) = \delta(g_t, g_{cc})$.

Nodes in $\mathcal{G}_t.V$ would have had multiple edges between them, as in Section II-C1. The molecular interaction network is simplified, to get ordinary graphs, using the following method.

The weight function $w_t : \mathcal{G}_t.V \times \mathcal{G}_t.V \rightarrow [0, +1]$ is defined as,

$$w_t(g_x, g_y) = \frac{|c(g_x, g_t)| + |c(g_y, g_t)|}{2} \quad (2)$$

where, g_x, g_y are *candidate causal genes* and $g_x \sim g_y$ (i.e., they have an interaction in any of the three ways outlined in Section II-C1) [1], [7].

Thus, multiple edges defined in Section II-C1 get substituted with a single weighted edge in \mathcal{G}_t .

The genes at the endpoints of edges with higher edge weights [1], [7] are said to have a role in the disease, or are *causal genes*.

The weight of a path p from g_t to $g_{cc} = \sum_{e \in g_t \xrightarrow{p} g_{cc}} w_t(e)$.

The paths from g_t to *candidate causal genes* having high weight values are called *dysregulated pathways*. There may be multiple *dysregulated pathways* in any (\mathcal{G}_t, w_t) . It may be recalled that there are $|G_t|$ such (\mathcal{G}_t, w_t) , or interaction networks, and each typically contains multiple *dysregulated pathways*.

Genes corresponding to nodes belonging to the *dysregulated pathways* are called the *causal genes*. i.e., $g_{cc} \in G_c$, if there exists g_t and a path p such that $g_t \xrightarrow{p} v$ and $g_{cc} \in p$ where, v is a *candidate causal gene*, g_t is a *target gene* and $g_t \xrightarrow{p} v$ is a *dysregulated pathway* in (\mathcal{G}_t, w_t) .

In the above description, the terms “high” edge weights and “high” weighted paths have been used. As these adjectives are abstract, they need to be quantified.

E. Towards a new algorithm

The problem defined in Section II-D is explained and stated below, without abstract adjectives.

A protein network can be represented as an edge-weighted directed graph with each g_t as the source vertex and $w_t(u, v)$ as the edge-weight between nodes u and v defined previously.

Nodes represent genes or proteins and edges represent molecular interactions.

(\mathcal{G}_t, w_t) obtained for each $g_t \in G_t$ is further reduced, by eliminating “low” weighted edges.

“High” and “low” are decided based on a *threshold*, which is defined as follows.

In (\mathcal{G}_t, w_t) , the *threshold* τ is defined as

$$\tau = \frac{\sum_{e \in \mathcal{G}_t.E} w_t(e)}{|\mathcal{G}_t.E|} \quad (3)$$

The set of edges to be removed, R , is computed as follows.

$$R = \{e | e \in \mathcal{G}_t.E \wedge w_t(e) < \tau\} \quad (4)$$

The edges in R are then removed from (\mathcal{G}_t, w_t) [9].

The paths having high path weights (as defined earlier) from g_t to all *candidate causal genes*, in graphs for all $g_t \in G_t$ are the *dysregulated pathways* and all vertices belonging to *dysregulated pathways* are *causal genes*.

Given n graphs (\mathcal{G}_i, w_i) with source vertex $i \in G_t$ and $\mathcal{G}_i.V = i \cup G_{cc}$. Find G_c .

$$G_c = \cup_{\forall (\mathcal{G}_i, w_i)} \{g_{cc} | w_i(g_{cc}, g_i) > \tau \\ \wedge \exists p : g_i \xrightarrow{p} g_{cc} \wedge \forall e \in p \quad w_i(e) > \tau\} \quad (5)$$

where, τ is the *threshold* value defined in Equation 3.

Find *dysregulated pathways* D_p as all the paths p satisfying the above condition. i.e.,

$$D_p = \{p | \forall e \in p \quad w_i(e) > \tau\} \quad (6)$$

It is inferred that every research work described in Section I attempts to find the complete set G_c and all *dysregulated pathways*. However, for realistic biological networks, the size of the sets and the number of paths is very large for practical computation [1], [7]. In order to get the results in reasonable computation time, the following heuristic is used in this work. “In a biological network, higher the *degree* of a node, more relevant it is” [31], [32], where *degree* of a node is the number of its neighboring edges. *Degree* of a node, v is denoted by $deg(v)$.

Hence, each time a path $p \in D_p$ is explored in (\mathcal{G}_i, w_i) , a vertex g_{cc} with $deg(g_{cc}) = \delta(\mathcal{G}_i, w_i)$ is removed. $\delta(\mathcal{G}_i, w_i)$ is the minimum degree of (\mathcal{G}_i, w_i) . Let v_δ be the node with minimum degree in (\mathcal{G}_i, w_i) . Then, (\mathcal{G}_i, w_i) is updated such that $\mathcal{G}_i.V = \mathcal{G}_i.V \setminus v_\delta$. Next, a path, $p \in D_p$ is traversed in (\mathcal{G}_i, w_i) and repeat this process till $|V| = k$, where k is an integer.

The process of removing v_δ can be made efficient in terms of execution time by mapping G_{cc} along with the degree of each node, $deg(g_{cc})$ to *min-heap* data structure. *Min-heap* is the underlying data structure of priority queues with the property of $A[parent(u)] \leq A[u]$, where u is a node other than root node and $parent(u)$ is the parent node of u [33]. Here, priority is given in terms of the decreasing order of degree. Highest priority is assigned to the vertex v_δ to ensure

that the vertex with the least degree is removed. Considering each node and its degree in the molecular interaction network as the elements of the *min-heap*, v_δ is deleted by removing the root node of the *min-heap*. Then, this data structure is reorganized to bring the highest priority node or v_δ at the root [33] and the process specified earlier continues till the number of nodes $= k$.

Each time v_δ is removed, its adjacent edges are also removed, thereby reducing the complex nature of biological networks which in turn results in identification of relevant genes and corresponding pathways within practical time interval.

Now, definitions of G_c and D_p are updated as follows:

$$G_c = \cup_{\forall (\mathcal{G}_i, w_i)} \{g_{cc} | w_i(g_{cc}, g_i) > \tau \\ \wedge \exists p : g_i \xrightarrow{p} g_{cc} \wedge \forall e \in p \quad w_i(e) > \tau \\ \wedge \exists g_{cc} : deg(g_{cc}) > \delta(\mathcal{G}_i, w_i)\} \quad (7)$$

Here, an additional characteristic of *causal genes* is incorporated apart from the attributes described in Equation 5: *causal genes* tend to be higher *degree* nodes.

Dysregulated pathways D_p as all the paths p satisfying the mentioned conditions. i.e.,

$$D_p = \{p | \forall e \in p \quad w_i(e) > \tau \wedge \exists v \in p : deg(v) > deg(v_\delta)\} \quad (8)$$

Finally, motivated by the approach in [9], *randomized rounding*, an efficient approach to design an *approximation algorithm* is done to fetch the most relevant paths in terms of weight, $w_i(e)$ and degree, $deg(g_{cc})$ in (\mathcal{G}_i, w_i) . The randomized rounding approach resulted in a factor $\frac{\psi}{\eta\kappa}$ algorithm. Here, ψ is the total number of paths in Ψ , where, Ψ is the collection of all distinct paths in the reduced network. η is the total number of distinct edges in Ψ and κ is the number of occurrences of a randomly selected edge in Ψ .

Its underlying principle follows. $\sum_{\forall e \in p} w_i(e)$ is calculated for (\mathcal{G}_i, w_i) , where $p \in D_p$. Let μ_p be $max(\sum_{\forall e \in p} w_i(e))$ in (\mathcal{G}_i, w_i) . *Dysregulated pathway* is taken as the path having value μ_p and the members of the pathway is taken as the *causal genes* for (\mathcal{G}_i, w_i) . Therefore, definitions of G_c and D_p are reformed by adding the following attribute: *causal genes* and thereby *dysregulated pathways* are constituted of maximum weighted paths after *randomized rounding*. Final definitions of G_c and D_p after joining all the mentioned attributes together are as given below.

$$G_c = \cup_{\forall (\mathcal{G}_i, w_i)} \{g_{cc} | w_i(g_{cc}, g_i) > \tau \\ \wedge \exists p : g_i \xrightarrow{p} g_{cc} \wedge \forall e \in p \quad w_i(e) > \tau \\ \wedge \exists g_{cc} : deg(g_{cc}) > \delta(\mathcal{G}_i, w_i) \\ \wedge \exists p : \forall e \in p \quad \sum w_i(e) = \mu_p\} \quad (9)$$

Dysregulated pathways D_p as all the paths p satisfying the

stated conditions. i.e.,

$$D_p = \{p | \forall e \in p \quad w_i(e) > \tau \wedge \exists v \in p : \deg(v) > \deg(v_\delta) \\ \wedge \forall e \in p \quad \sum w_i(e) = \mu_p\} \quad (10)$$

F. Implementation and Validation of the results

Comparative study of the related works in literature and the proposed method has been performed based on the accuracy of results and the execution time. All these methods have been implemented in Intel @Core i7-3770 CPU @ 3.40GHz X 8 machine with memory (RAM) capacity of 16 GB by using the programming languages R and C. Also, the effectiveness of the Algorithm 1 has been explored by using Reactome Pathway Database to provide its biological significance.

1) *Analysis based on τ and k* : As the threshold defined in 3, is quite critical for removing the low-weighted edges, its effect on identifying causal genes and dysregulated pathways has been simulated. For this, apart from implementing Algorithm 1, an algorithm without using steps 4, 5 and 6 in Algorithm 1 has been implemented. Also, Algorithm 1 has been implemented against different values of k which is defined in Section II-E for the datasets.

III. DISCUSSION AND RESULTS

A. Approximation algorithm with graph reduction

Pseudo code for the algorithm to identify causal genes and dysregulated pathways is given in Algorithm 1. The algorithm considers a molecular interaction network with multiple sources and sinks. Source set is the set of target genes and sink set is the set of candidate causal genes other than target genes. Proof of the approximation factor is given in Appendix ??.

B. Implementation and Validation of the results

All the related works in literature and the proposed algorithm have been implemented as specified in Section II-A on multiple datasets. For Breast cancer dataset, G_t consists of 41 genes and G_{cc} consists of 309 genes, for Lung cancer dataset, G_t consists of 7 genes and G_{cc} consists of 70 genes and for Pancreatic cancer dataset, G_t consists of 46 genes and G_{cc} consists of 140 genes. (G_t, w_t) have 8403 edges and 350 nodes with a total of 41 such graphs for Breast cancer dataset, 466 edges and 77 nodes with a total of 7 such graphs for Lung cancer dataset and 729 edges and 186 nodes with a total of 46 such graphs for Pancreatic cancer dataset. Benchmark data as specified in Section II-B have been obtained for the cases Breast cancer, Lung cancer and Pancreatic cancer as 26 genes, 21 genes and 83 genes respectively.

Let us denote Breast cancer dataset as dataset I, Lung cancer dataset as dataset II and Pancreatic cancer dataset as dataset III.

1) *Simulation study based on τ* : The effect of τ on identifying causal genes and dysregulated pathways has been simulated by considering multiple graphs with *SLCO2A1* as source for a subset of sink nodes on dataset I, *Cdk6* as source for a subset of sink nodes on dataset II and *Acaca* as source

Algorithm 1 Infer causal genes and dysregulated pathways

Input: n graphs (G_i, w_i) with source vertex $i \in G_t$ and sink vertex $g_{cc} \in G_{cc}$

Output: G_c and D_p

```

1: for all  $i \in G_t$  do
2:   for all  $g_{cc} \in G_{cc}$  do
3:      $s \leftarrow g_{cc}$ 
4:     Calculate  $\tau$ 
5:     Find  $R$ 
6:     Remove  $R$  from  $(G_i, w_i)$ 
7:     repeat
8:        $\Psi = \Psi \cup \text{Findpath}((G_i, w_i), i)$ 
9:     until  $(|V| = k)$ 
10:     $\psi \leftarrow |\Psi|$ 
11:    if  $\psi = 0$  then
12:      goto step 1
13:    else if  $\psi = 1$  then
14:       $D_p = \Psi$ 
15:      Find  $G_c$ 
16:      goto step 1
17:    else
18:      Find  $\eta$ , total number of distinct edges in  $\Psi$ 
19:      Randomly pick an edge,  $e$  from a path in  $\Psi$ 
20:      Calculate  $\kappa$ , number of occurrences of  $e$  in  $\Psi$ 
21:       $r \leftarrow \frac{\eta\kappa}{\psi}$ 
22:      for  $i = 1$  to  $r$  do
23:        Randomly select a path  $p$  in  $\Psi$ 
24:        Calculate  $\sum_{e \in p} w_i(e)$ 
25:      end for
26:       $D_p \leftarrow \text{Path with value } \mu_p$ 
27:      Find  $G_c$ 
28:    end if
29:  end for
30: end for

```

for a subset of sink nodes on dataset III, as specified in Section II-F1. The results are summarized and tabulated in Table II.

Dataset	# Causal Genes	Execution Time
I	[86][87]	[435.35 Sec.][737.33 Sec.]
II	[24][35]	[1.503 Sec.][1.893 Sec.]
III	[8][9]	[45.5 Sec.][800.119 Sec.]

TABLE II: Effect of τ with the order [with τ][without τ]

There is a slight increase in the number of causal genes identified in the method without using τ ; but, it compromises on execution time while considering all the genes.

2) *Analysis based on k* : Algorithm 1 has been implemented against different values of k as defined in Section II-E for the datasets. Considering the size of the datasets, for dataset I, a subgraph of (G_t, w_t) has been considered by selecting first 100 genes after sorting p -value, resulting in 8 graphs of 432 edges and 74 nodes with benchmark data consisting of

```

procedure Findpath( $(\mathcal{G}_i, w_i), v$ )
  if  $v = s$  then
     $p = p \cup v$ 
     $\Psi \leftarrow p$ 
     $\mathcal{G}_1.V = \mathcal{G}_i.V \setminus \{i, s\}$ 
    Calculate  $deg(g_{cc})$  for all  $g_{cc} \in \mathcal{G}_1.V$ 
    Assign priority based on decreasing order of degree
    Make a min-heap for  $(\mathcal{G}_1)$ 
    Find  $v_\delta$ 
    Delete  $v_\delta$  from min-heap
    Update  $\mathcal{G}_i.V = \mathcal{G}_i.V \setminus v_\delta$ 
  else
    for all neighbours  $adj$  of  $v$  do
      if  $adj$  is unexplored then
         $p = p \cup v$ 
        Findpath( $(\mathcal{G}_i, w_i), adj$ )
      end if
    end for
  end if
end procedure

```

10 genes. The receiver-operating characteristic (ROC) analysis has been used to compare the performance of the algorithm under various values of k . The values of k are selected as 10% of $|V|$, 20% of $|V|$, 30% of $|V|$ and 40% of $|V|$. The ROC curve plots the true positive rate (TPR) / sensitivity versus the false positive rate (FPR) / (100-specificity) where, TPR is the percentage of causal genes which are correctly identified based on benchmark dataset and FPR is the percentage of causal genes which are not present in the benchmark dataset. To compare different curves obtained by ROC analysis, the area under curve (AUC) for each curve has been calculated, which is given in Table III. Higher the value of AUC, better the result is. Also, the execution time for each case is taken and it is summarized in Table IV. When the ROC curve and execution time are compared, $k=10\%$ of $|V|$ has shown better performance and it has been selected as the cut-off value in Algorithm 1 provided overall accuracy is same in all cases.

Values of k	Dataset I	Dataset II	Dataset III
10% of $ V $	0.0294	0.6932	0.656
20% of $ V $	0.0294	0.5263	0.694
30% of $ V $	0.0294	0.3982	0.579
40% of $ V $	0.0294	0.4463	0.552

TABLE III: AUC for different values of k

Values of k	Dataset I	Dataset II	Dataset III
10% of $ V $	0.16	6.68	90.291
20% of $ V $	1.5	9.57	148.28
30% of $ V $	10.9	10.38	167.206
40% of $ V $	14.06	14.24	1637

TABLE IV: Execution time (Seconds) for different values of k

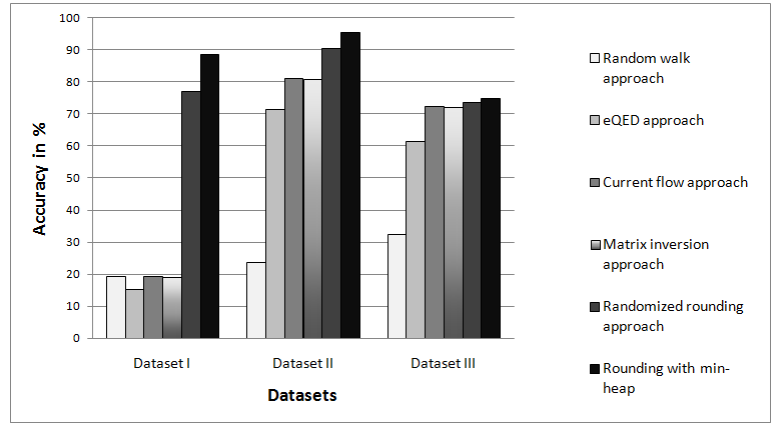


Fig. 1: Comparison based on the accuracy of results

C. Measuring the accuracy of results and Analysis based on the execution time

Accuracy is defined as the measurement of closeness between the benchmark data and the actual results. It is calculated as the percentage of causal genes which are correctly identified based on benchmark dataset. Comparison based on the accuracy of results is given in Figure 1 and the execution time is given in Table V.

Approaches	Dataset I	Dataset II	Dataset III
Random walk approach	65 (Min.)	5 (Sec.)	1 (Min.)
eQED approach	33 (Min.)	4 (Sec.)	0.8 (Min.)
Current flow approach	37 (Min.)	1801 (Hours)	1226 (Min.)
Matrix inversion approach	34 (Min.)	1789 (Hours)	1212 (Min.)
Randomized rounding approach	4114 (Min.)	53 (Hours)	1.73 (Min.)
Rounding with min-heap	2155 (Min.)	8.61 (Sec.)	1.12 (Min.)

TABLE V: Comparison based on the execution time

Observations made based on these results are given next. In random walk approach, the Pearson's correlation coefficient of the *gene expression levels* of genes at each node and the *target gene* is calculated and transition probability has obtained using this value. Random walk based on transition probabilities is done multiple times. *Disease causing genes* including their *pathways* are obtained as the result. Random walk approach takes less amount of execution time, but the number of genes identified and the accuracy of results are less. The molecular interaction is considered as an electric circuit according to other approaches in literature. The conductance of each link is calculated based on the Pearson's correlation coefficient of the *gene expression levels* of genes at each node and *target gene*. The Kirchhoff's current law and Ohm's law are utilized to obtain voltages. The current is calculated using Ohm's law for each edge in the network. *Disease causing genes* and *dys-regulated pathways* are identified using each of these methods. eQED approach takes least amount of execution time, but gives lesser number of genes and low accuracy compared to methods other than random walk approach. Electric circuit approach with multiple sources and sinks and fast iterative

matrix inversion return same number of disease causing genes with same accuracy. Here, genes receiving current of at least 70% of the maximum current among all genes are considered in each iteration and thereby its execution time is greater in most of the cases. Since, fast iterative matrix inversion uses fourth order iterative method, execution time is less compared to electric circuit approach with multiple sources and sinks. Approximation algorithm outperforms other methods in terms of the number of disease causing genes identified and the accuracy of results. But, it takes much more time compared to random walk approach and eQED approach. Here, before applying randomized rounding, all possible paths are identified which in turn results in increased execution time.

The proposed algorithm, Approximation algorithm with graph reduction has been implemented on multiple datasets and the sets of *causal genes* and associated *pathways* for the disease cases have been identified. Initially, a set of less-weighted edges are removed. Then, a set of vertices are removed together with the exploration of a set of distinct simple paths by making sure that a vertex of least degree is deleted. Min-heap data structure is utilized for time efficient vertex removal. This process repeats k times with its value as 10% of $|V|$. Finally, most relevant paths are identified using randomized rounding. This newly proposed approximation algorithm with graph reduction defeats all other methods by identifying more number of genes within less time interval.

The sets of all newly identified *disease causing genes* apart from the known genes in the benchmark datasets as well as the resultant sets of identified *causal genes* and *dysregulated pathways* for all the datasets are given in the Supplementary material.

IV. CONCLUSION AND FUTURE SCOPE

Computationally intensive process of identifying causal genes and related pathways from the huge molecular interaction networks has been addressed in this paper. The huge size of the molecular interaction network is reduced by retaining more relevant nodes and edges in terms of edge-weights and connectivity. Then, most appropriate nodes, thereby paths in terms of these parameters are selected by using the concept of *approximation*. Experimentations proved that the newly proposed approximation algorithm with graph reduction outperforms all other methods by identifying more number of genes within lesser time.

REFERENCES

- [1] S. W. Yoo-Ah Kim, Jozef H Przytycki and T. M. Przytycka, "Modeling information flow in biological networks," in *Physical Biology*. 8(2011) 035012(9pp). 2011 IOP Publishing Ltd, 13 May 2011, pp. 1–9.
- [2] S. Winslow, K. Leandersson, A. Edsjö, and C. Larsson, "Prognostic stromal gene signatures in breast cancer," *Breast Cancer Research*, vol. 17, no. 1, pp. 1–13, 2015.
- [3] P. T. Cho D-Y, Kim Y-A, "Chapter 5: Network biology approach to complex diseases," *PLoS Computational Biology*, vol. 8(12), 2012.
- [4] Z. Tu, L. Wang, M. N. Arbeitman, T. Chen, and F. Sun, "An integrative approach for causal gene identification and gene regulatory pathway inference," *Bioinformatics*, vol. 22, no. 14, pp. e489–e496, 2006.
- [5] P. G. Doyle and J. L. Snell, *Random Walks and Electric Networks*. Washington DC: Mathematical Association of America, 2000.
- [6] S. Suthram, R. M. K. Andreas Beyer, and T. I. Yonina Eldar, "eqed: an efficient method for interpreting eqtl associations using protein networks," *Molecular Systems Biology*, vol. 4:162, 2008.
- [7] P. T. Kim Y-A, Wuchty S, "Identifying causal genes and dysregulated pathways in complex diseases," *PLOS Computational Biology*, vol. 7, 2011.
- [8] T. Jose and P. Chandran, "A fast iterative method for modeling information flow in biological networks," Master's thesis, National Institute of Technology Calicut, India, 2013.
- [9] J. V. Devasia and P. Chandran, "Towards an improved algorithm for modeling information flow in biological networks," in *Interanational Conference on Advances in Computing, Communications, and Information Science*. Elsevier, 2014, pp. 88–95.
- [10] R. A. Irizarry, B. Hobbs, Collin *et al.*, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003.
- [11] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez gene: gene-centered information at ncbi," *Nucleic Acids Research*, vol. 33, no. suppl 1, pp. D54–D58, 2005.
- [12] S. A. Forbes, D. Beare, P. Gunasekaran, K. Leung, N. Bindal, H. Boutselakis, M. Ding, S. Bamford, C. Cole, S. Ward, C. Y. Kok, M. Jia, T. De, J. W. Teague, M. R. Stratton, U. McDermott, and P. J. Campbell, "Cosmic: exploring the world's knowledge of somatic mutations in human cancer," *Nucleic Acids Research*, vol. 43, no. D1, pp. D805–D811, 2015.
- [13] M. J. C. Sandrine Dudoit, Yee Hwa Yang and T. P. Speed, "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments," *Statistica Sinica*, vol. 12, pp. 111–139, 2002.
- [14] E. S. Kawasaki, "The end of the microarray tower of babel: Will universal standards lead the way?" *Journal of Biomolecular Techniques*, vol. 17:200206, 2010.
- [15] F. Millenaar, J. Okyere, S. May, M. van Zanten, L. Voesenek, and A. Peeters, "How to decide? different methods of calculating gene expression from short oligonucleotide array data will give different results," *BMC Bioinformatics*, vol. 7, no. 1, p. 137, 2006.
- [16] J. W. Ho, M. Stefani, C. G. dos Remedios, and M. A. Charleston, "Differential variability analysis of gene expression and its application to human diseases," *Bioinformatics*, vol. 24, no. 13, pp. i390–i398, 2008.
- [17] M. I. "Epistasis: Gene interaction and phenotype effects," *Nature Education*, vol. 1(1):197, 2008.
- [18] G. stlund, M. Lindskog, and E. L. L. Sonnhhammer, "Network-based identification of novel cancer genes," *Molecular & Cellular Proteomics*, vol. 9, no. 4, pp. 648–655, 2010.
- [19] R. Kelley and T. Ideker, "Systematic interpretation of genetic interactions using protein networks," *Nat Biotechnology*, vol. 23(5), 2005.
- [20] C. von Mering, L. J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Krger, B. Snel, and P. Bork, "String 7 recent developments in the integration and prediction of protein interactions," *Nucleic Acids Research*, vol. 35, no. suppl 1, pp. D358–D362, 2007.
- [21] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguéz, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. v. Mering, "The string database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Research*, 2010.
- [22] A. Shojaie, "Link prediction in biological networks using multi-mode exponential random graph models," in *Eleventh Workshop on Mining and Learning with Graphs, ACM 978-1-4503-2322-2*. ACM, 2013.
- [23] A. Ferro, R. Giugno, G. Pigola, A. Pulvirenti, D. Skripin, G. D. Bader, and D. Shasha, "Netmatch: a cytoscape plugin for searching biological networks," *Bioinformatics*, vol. 23, no. 7, pp. 910–912, 2007.
- [24] G. Pavlopoulos, M. Secrier, C. Moschopoulos, T. Soldatos, S. Kossida, J. Aerts, R. Schneider, and P. Bagos, "Using graph theory to analyze biological networks," *BioData Mining*, vol. 4, no. 1, p. 10, 2011.
- [25] Y.-A. Kim, S. Wuchty, and T. M. Przytycka, "Simultaneous identification of causal genes and dys-regulated pathways in complex diseases," in *Proceedings of the 14th Annual international conference on Research in Computational Molecular Biology*, ser. RECOMB'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 263–280.
- [26] O. Odibat and C. K. Reddy, "Ranking differential hubs in gene co-expression networks," *Journal of Bioinformatics and Computational Biology*, vol. 10, p. 1240002 (15 pages), 2012.
- [27] A. Devonshire, R. Elasarapu, and C. Foy, "Evaluation of external rna controls for the standardisation of gene expression biomarker measurements," *BMC Genomics*, vol. 11, no. 1, p. 662, 2010.

- [28] C. Wu, J. Zhu, and X. Zhang, "Network-based differential gene expression analysis suggests cell cycle related genes regulated by e2f1 underlie the molecular difference between smoker and non-smoker lung adenocarcinoma," *BMC Bioinformatics*, vol. 14, no. 1, p. 365, 2013.
- [29] A. C. A. Fredrik Barrenäs, Sreenivas Chavali *et al.*, "Highly interconnected genes in disease-specific networks are enriched for disease-associated polymorphisms," *Genome Biology*, 2012.
- [30] M. A. P. Sumeet Agarwal, Charlotte M. Deane and N. S. Jone, "Revisiting date and party hubs: Novel approaches to role assignment in protein interaction networks," *PLOS Computational Biology*, vol. 6(6), p. e1000817, 2010.
- [31] P. F. Jonsson and P. A. Bates, "Global topological features of cancer proteins in the human interactome," *Bioinformatics*, vol. 22, no. 18, pp. 2291–2297, 2006.
- [32] G. K-I, V. D. Cusick ME, V. M. Childs B, and B. A-L, "The human disease network," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 21, pp. 8685–8690, 2007.
- [33] C. E. Leiserson, C. S. Thomas H. Cormen, and R. Rivest, *Introduction to Algorithms*. MIT Press, 2001.